BU-1336-M

DERIVATION OF THE PROBABILITY MASS FUNCTION OF A PURE BIRTH PROCESS FROM A DEATH PROCESS.

by

Carlos M. Hernández-Suárez

Carlos M. Hernández-Suárez is a graduate student at the Biometrics Unit, Cornell University, Ithaca, NY 14853-7801. Internet address cmh1@cornell.edu. This work was partially supported by a Conacyt Scholarship (Mexico) and by grant DEB-9253570 (Presidential Faculty Award) to Carlos Castillo-Chávez. The author thanks George Casella, Olga Cordero and Lynn Eberly for their comments.

1. Abstract

For a population that grows according to the rules of a pure birth process, the probability that it will reach a certain size by time t is usually derived by solving a system of differential equations. Students may find an alternative derivation based on elementary statistical properties useful. We present a solution that relates a pure birth process with a pure death process. Although the concept of integration is required, no actual integration is needed. Familiarity with some basic properties of the geometric, exponential, binomial and negative binomial distributions is assumed.

Keywords : *Birth death process*, *Yule process*, *Exponential Distribution.*

2. Introduction

The pure birth process (PBP) was first proposed by Yule (1925) to describe the rate of evolution of a new species within a genus. In this process, every individual (species) acts independently of the others and gives birth to a new individual at a random time which follows the exponential distribution with parameter λ. We can assume that individuals do not die or that each birth consists of a splitting of an individual. The population size at time t is the number of births up to time t plus the initial population size n_0. It is also assumed that births occur instantaneously and that offspring are ready to reproduce immediately. Although the set of assumptions is unrealistic, it is the simplest stochastic process which involves reproduction by "splitting" (see Renshaw,1991).

3. Methodology

3.1 Initial Population size equal to 1

Consider first a population that starts with one individual. In order to compute the probability that the population size at time t will be n in a PBP, we must calculate the probability of having $n-1$ births before time t. In contrast with the pure death process (PDP) case, calculating this probability is rather difficult, since in a PDP, deaths are independent and the probability that a single individual is alive at time t is $n\,\theta^{n-1}(1-\theta)$, where θ is the probability that an individual dies before time t. Moreover, in a PBP, the number of individuals by time t could be in theory any positive integer, and we think of events as not so independent, since the i th individual must have born before the $i+1$st. It

will be shown that this is a fallacy and that the logic behind the two processes is very simple.

Let $P_{X:Y}(t)$ denote the probability that the population becomes of size Y by time t starting from size X. If $X < Y$ this is a PBP and conversely if $X > Y$ we have a PDP. Notice that "by time t" implies that the population needes to became of size Y at some time t' prior to t and remained that size during $t - t'$. Define also $f_{X:Y}(t)$ as the density function "time to reach size Y starting from X". We can write $P_{1:n}(t)$ as

$$P_{1:n}(t) = \int_0^t \int_{t_1}^t f_{1:2}(t_1)\ f_{2:n}(t_2 - t_1)\ e^{-n\lambda(t-t_2)} dt_2 dt_1$$

where the last term inside the integral is the probability that once the process reached size n at time t_2, no births occured in $(t - t_2)$. Notice $f_{1:2}(t_1)$ is the pdf of an exponential random variable with parameter λ, therefore we can rewrite the last expression as follows:

$$P_{1:n}(t) = \int_0^t \int_{t_1}^t \lambda e^{-\lambda t_1}\ f_{2:n}(t_2 - t_1)\ e^{-n\lambda(t-t_2)} dt_2 dt_1 \qquad (1)$$

On the other hand, $P_{n:1}(t)$ can be expressed as:

$$P_{n:1}(t) = \int_0^t \int_{t_1}^t f_{n:n-1}(t_1)\ f_{n-1:1}(t_2 - t_1)\ e^{-\lambda(t-t_2)} dt_2 dt_1$$

where the last term reflects the fact that no deaths occur once the population reaches size 1. Since $f_{n:n-1}(t_1)$ is the pdf of an exponential random variable with parameter $n\lambda$, the last expression can be rewritten as:

$$P_{n:1}(t) = n \int_0^t \int_{t_1}^t \lambda e^{-n\lambda(t_1)}\ f_{n-1:1}(t_2 - t_1)\ e^{-\lambda(t-t_2)} dt_2 dt_1 \qquad (2)$$

Now, notice that $f_{2:n}(t)$ is the pdf of the sum of $n-2$ exponential random variables whose rates are 2,3,4,...,$n-1$, and similarly $f_{n-1:1}(t)$ is the pdf of the sum of $n-2$ random variables whose rates are $n-1$, $n-2$,...,2. It turns out that

$$f_{2:n}(t) = f_{n-1:1}(t)$$

In (2) apply the change variable $t_1 = t - t_2$ and $t_2 = t - t_1$ and we get:

$$P_{n:1}(t) = n \int_0^t \int_{t_1}^t \lambda e^{-n\lambda(t-t_2)} \; f_{n-1:1}(t_2 - t_1) \; e^{-\lambda(t_1)} dt_2 dt_1$$

that is

$$P_{n:1}(t) = n \, P_{1:n}(t) \tag{3}$$

Having developed a simple relationship between the processes, we now derive the formula for $P_{n:1}(t)$.In a PDP, the probability that a given individual is alive by time t is:

$$1 - F(t) \; = \; e^{-\lambda t}$$

Since deaths are independent events, the number of individuals alive by time t is a binomial random variable with parameters n and $e^{-\lambda t}$, then

$$\begin{aligned} P_{n:1}(t) &= \binom{n}{1} e^{\lambda t} \left(1 - e^{-\lambda t}\right)^{n-1} \\ &= n e^{-\lambda t} \left(1 - e^{-\lambda t}\right)^{n-1} \end{aligned}$$

In view of (3), we finally arrive to an expression for $P_{1:n}(t)$:

$$P_{1:n}(t) = e^{-\lambda t} \left(1 - e^{-\lambda t}\right)^{n-1} \tag{5}$$

3.2 General Expressions

Now we use the properties of some well known distributions to generalize (5) to an initial population size n_0. From (5) we can see that the number of individuals at time t in a PBP that starts with a single individual follows a geometric distribution with parameter $e^{-\lambda t}$, therefore, the number of individuals at time t starting from n_0 is the sum of n_0 of these random variables, which is a negative binomial random variable with parameters n_0 and $e^{-\lambda t}$. Hence:

$$P_{n_0:n}(t) = \binom{n-1}{n_0-1} e^{-\lambda t}\left(1-e^{-\lambda t}\right)^{n-n_0} \quad n_0 < n$$

On the other hand, (4) can be generalized to any k in $[0, n]$ since the number of individuals alive is binomial; thus

$$P_{n:k}(t) = \binom{n}{k} \left(e^{-\lambda t}\right)^{k} \left(1-e^{-\lambda t}\right)^{n-k} \quad k < n$$

4. References

Renshaw, E. (1991) "Modelling Biological Populations in Space and Time". *Cambridge University Press*. p. 15-44.

Yule,G.U. (1925)"A Mathematical Theory of Evolution", *Philosophical Transactions of the Royal Society of London*. (B, 213), 21-87.